一个基于 RSV 索引检索技术的中英文 WEB 搜索引擎

余晓 丁伟 魏星　　Peter Schauble
（东南大学计算机科学与工程系 南京 210096）（瑞士苏黎世联邦工业大学信息学院）
xyu@njnet.edu.cn　　schauble@inf.ethz.ch

【摘要】本文介绍了 CERNET 华东（北）地区网络中心和 ETH 联合开发的一个基于 RSV 索引检索技术的中英文 WEB 搜索引擎－－网达系统，讲述了该系统的基本情况、结构和主要工作原理，以及系统的主要技术特点－－运行效率高、Feedback 功能、feature 的获取及配置文件，并与国内常用的搜索引擎通过实例作了比较。

【关键词】搜索引擎，信息检索，RSV，计算机网络

一. 引言

随着科技的不断发展，信息变得越来越轻量化， 越来越多的信息存放在万维网上，万维网就象一个大型的电子信息库，然而其中的信息与原来的信息相比具有了新的特点：分布广、无结构性、变化快等。如果能够借鉴原有成熟的信息检索技术的思想，利用它的技术，结合现代信息的新特点设计万维网上信息检索系统无疑将会是一种很好选择。

由 CERNET 华东（北）地区网络中心和 ETH（瑞士苏黎世联邦工业大学）联合开发的网达系统采用了经典的 RSV 索引检索技术为核心进行 Web 中英文搜索，目前的网址为 http://peony.njnet.edu.cn:8888/WangDa/（图 1）。

图 1 网达主页

二. 系统的基本情况

网达系统从 98 年 6 月上旬开始试运行，可接受以中英文输入的查询，内容包括了以 CERNET 为主的中国大陆各互联网的网页信息，目前该系统的信息采集部分每天定时工作，随着时间的推移，索引库将逐渐扩大，系统的覆盖范围和用户查询的满意程度也将随之逐步提高。

与普通的 WEB 信息查询系统一样，网达也通过提供一个输入框支持用户的查询。在用户向系

统提交了查询（query）后，系统将查询结果根据 RSV 值的高低降序排列，10 个文档为一页返回，并用“前页文档”和“后页文档”两个功能键支持用户对一次查询的结果进行翻阅。

从用户功能角度而言，与目前网上已有的信息查询系统相比，网达的主要特点是：

- 该系统采用信息检索的技术思想，不仅对 Spider 得到的文档进行自动的索引(index)处理，也对用户的查询词语进行自动的索引处理，因此具有宽大的输入区，用户无需使用“AND”、“OR”等逻辑连接词来组织自己的查询关键词，而只需用自然语言描述查询需求键入输入区即可。
- 对每个返回的文档均给出“加亮”(highlight) 和“直接”(direct link) 两条链接（图 2）。前者以结果文档的标题(title)为链接提示，选择它将获得由本引擎返回的该文档的全部正文信息，其中与用户查询有关的文字将亮化以区别于其它文字。后者以该文档的 URL 为链接提示，选择它将直接连往该文档所在的服务器。

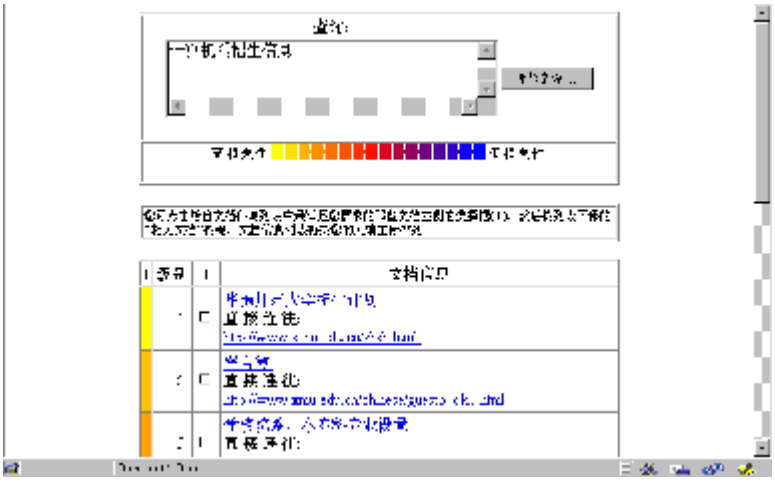

图 2 网达结果页面

- 通过“相关文档”的功能键向用户通过“用户回馈”(feedback)功能。Feedback 功能支持用户根据本次查询的目的，从自身的语义角度在查询结果中选择满意程度较高的文档，形成新的查询（query），对用户选定的这些文档进行相关文档的查询。 这样不仅可以使得查询结果的满意程度得到相应的提高，也可以使得用户在得到所需的信息之后进行相关内容的查询（或深度查询）。

三. 网达系统的结构和主要工作原理

网达系统的简单结构和工作流程如图 3 所示。

网达系统分为信息搜集服务器和 IR 服务器两个部分， 它们的读数据操作完全独立。在数据修改时，信息搜集服务器获得的数据和 IR 数据（即 IR index）通过 Synchronizer 达到一致性。信息搜集服务器顺序从 Start.html 中存放的 URL 开始，按照管理者定义的一定深度进行页面搜索。 IR Server 管理一个 IR index（即信息检索索引）的访问结构。

一旦有某个文档被插入、删除或修改，系统就会将一个元组插入一个内部的 update 表中，该元组包括文档的标识 $ID(d_j)$ 和 d_j 更新操作的有关信息。synchronizer 定期的读 update 表，如果在 update 表出现了新的内容，synchronizer 将会执行相应的操作。

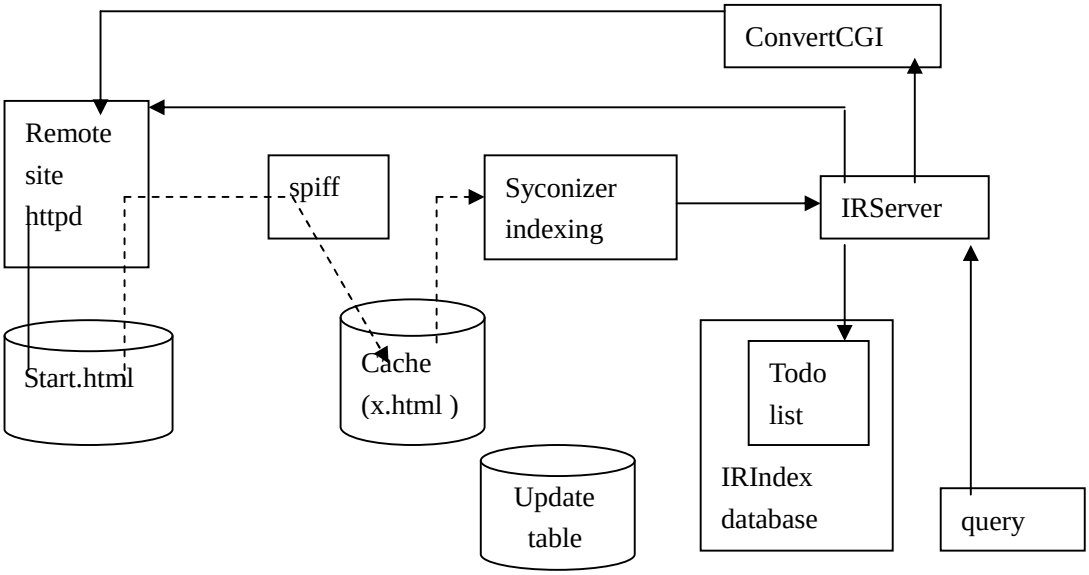

图 3 网达系统的简单结构和简明工作流程

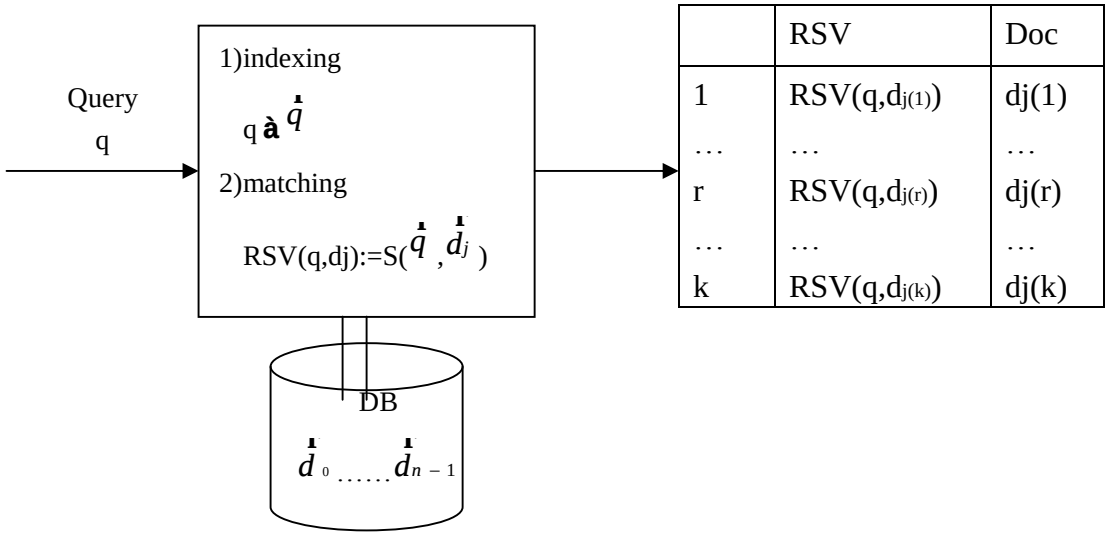

图 4 网达系统的查询过程

IR index 的访问结构决定于检索的方法（即存放什么样的 IR 数据）和检索的算法。下面的 IR 数据在网达系统中被存放：

$ff(j_i, d_j)$ feature j_i 在文档 d_j 中出现的频

$df(j_i)$ feature j_i 的文档频率，即包含 feature j_i 的文档数

网达系统采用了一种快速检索算法，基于对反向 posting list 的操作，即三元组

$$(ID(d_j), j_i, ff(j_i, d_j))$$

按 feature 存放而不是文档（document ID）。这样检索时将只访问那些在查询（query）中的 feature

的 posting 即可。同时，系统也存放一个正向的 posting list 以保证在删除文档时能够正确地将该文档的所有反向 posting list 删除，并且它也用于支持 feedback 功能。

用户查询后，系统给出两条链接，一条直接连往 remote site，另一条通过 ConvertCGI 将 remote site 的信息取回，并将文字作了处理，匹配词标为红色。

查询的具体流程如图 4 所示。

用户输入查询词 q，系统将查询语句按一定方法切开成 feature 向量 $\vec{q}$，计算 RSV(q, d_j)，最后按 RSV 的降序返回文档信息。

网达系统的 RSV 值计算符合以下原则：

1) 查询 q 中的 feature 在某个文档中出现的次数越多则该文档的 RSV 值越大；
2) 查询 q 中的 feature 在某个文档中出现的密度越高则该文档的 RSV 值越大；
3) 查询 q 中的某个 feature 在越多的文档中出现，则该 feature 对于 RSV 值的影响越小。

四. 网达系统的几个主要技术特点

1．系统的运行效率高

在网达系统中，经索引处理而获得系统所关心的参数(features 和相应的频数)后，cache 中的原始文档就没有存在的价值了，因此网达系统不以任何形式保留原始文档。网达系统的 IR index 数据库作为信息搜集部分和信息查询部分的交互点和整个系统的核心，其规模和代价要比保存原文少得多。Posting list 和反向 posting list 均按特定的方式排序，并使用随机查找函数方式进行搜索。由此可以看出网达系统对用户查询的处理、对硬盘空间的使用等方面的效率是非常高的，它可以在一台规模较小的主机上运行，并获得较为理想的效果；

2．Feedback

Feedback 是网达所特有的一项功能，国内 WEB 搜索引擎中还没有此方面的应用，国际上也不多见。由于本系统吸取了信息检索中的思想，采用了 RSV 索引检索方法，并且同时存储 posting list 和反向 posting list 两个表。两表的同时使用即保证了数据的一致性，又使得系统能够很快将用户选出的文档作为新的 query 重新查询成为了可能。

3．Feature 的获取

WEB 中文信息检索中 Feature 获取的问题其实本质上是中文信息处理中词切分的问题。网达采用了使用 Stopword （即指那些没有实际意义的常用词，如“我们”等）表的无词库的 overlapping bi-gram 的方法。系统将每一段中文文字滤去其中的 stopword 后，对剩下的各部分按从前到后的顺序每两个字取作一个 feature，如：古代历史，则取作“古代”“代历”“历史”三个 feature。这种处理方法有如下几个优点：

1) 中文词大部分是两个字为一词，所以大部分的词可以以词的形式成为 feature。
2) 连续的两个词之间相连的两个字在很多情况下仍有提高准确率的作用。如上例中的“代历”，在其它 feature（“古代”、“历史”）出现次数相同的情况下，可以使含有“古代历史”的文档相关性高于不含有“代历”字样的文档。
3) 不必建立强大的中文词库。

但这种处理方法没有考虑中文语义方面的问题，存在着一定的缺陷。

4．配置文件

网达系统的管理灵活，大多数管理可以通过修改它的配置文件实现。在目前的网达系统中，从信息搜集部分的定时启动、搜集范围的确定，到查询返回页面的说明文字、按键都可通过配置文件方便地调整，这使得系统的升级和修改都能比较方便地进行。

五. 实例比较

国内目前 CERNET 上已有多个著名的 WEB 搜索引擎，下面给出在两个具体的实例下它们与网达的运行情况。

【实例 1】Query: 江苏地理

- 北大天网系统的查询情况是：在第一页返回的 10 个文档条目中，准确匹配的有 4 个，其它的也都有一定联系；
- 清华指南针系统的查询情况是：只返回 2 个文档，均不匹配。若将 query 改为“江苏 地理”，（在江苏和地理间增加一个空格，表示是两个搜索词）则返回第一页 10 个条目中，准确匹配 3 个，其它的联系不大；
- 网达系统的结果是：在第一页返回的 10 条中，准确匹配的有 2 个，有一定联系的 1 个，其它的联系不大，若使用 feedback，选中准确匹配的 2 项的选择框后，按下“相关文档”键，重新返回的文档第一页准确匹配 3 条。

【实例 2】Query：计算机系招生信息

- 天网系统第一页准确匹配 3 个，第二页准确匹配 1 条；
- 指南针若直接原文键入，返回 2 个准确匹配的条目，若键入“计算机系 招生 信息”，第一页准确匹配 3 个，第二页 1 个；
- 网达第一页准确匹配 5 个，第二页 3 个，若使用 feedback，选中第一页准确匹配的 5 个选择框后，按下“相关文档”键，重新排列后返回的文档第一页 10 条全部准确匹配，第二页条准确匹配 4 条。

需要说明的是每个搜索引擎的查询结果均与其自身数据库中保存信息的数量和类别等密切相关，在一般情况下，数据库中的信息是动态的，每天都会发生变化。网达系统由于运行时间较短，因此库存的信息量目前还非常有限。上述两个试验的完成时间是 98 年 6 月中旬。

六. 结论

信息检索专家认为，信息检索工作的实质是用一切可能的方法和手段，在海量的文档信息和其使用者之间建立最有效的连接。在此基础上，多语言信息访问领域的专家更是认为高度抽象的信息检索方法应该是能够跨越语言边界的。网达系统实际上是将一种在西文检索方式中行之有效的方法用之于中文检索，我们认为这是一个有意义的尝试。事实上，评价一个搜索引擎是可以有许多角度的。网达的优点是系统的工作原理和设计思想简单高效，可以在各种类型的主机上运行。它的缺点主要是在对中文的语义处理方面，feature 的产生方法还不十分合理。实际上这个问题也是中文信息处理领域的所共同面临的一个问题。

目前网达系统安装在一台 Sun Sparc20 上， 还属于试运行阶段，由于主机能力有限，同时还由于系统运行时间较短积攒的信息量还相当较少，对于某些查询的结果可能尚不够理想，我们将在今后对系统进行进一步的改进后将其移植到较大规模的服务器上。

六．参考文献

[1] Peter Schauble, “Multimedia Information Retrieval”,1997，U.S.A., Kluwer Academic Publishers

[2] Gerard Salton and Michael J. McGill, “Introduction to Modern Information Retrieval”，1983，Japan，McGraw-Hill Book Company

[3] Jean-Paul Ballerini and Marco Buchel, etc. “Spider Retrieval System at TREC-5”, http://trec.nist.gov/pubs/trec5/t5_proceedings.html

A Chinese-English WEB Search Engine of RSV Indexing Retrieval Technique

【Abstract】 A cooperatively developed Chinese-English WEB search engine named Wangda is introduced in this paper. It works on the principle of RSV and has the highlight links and user’s feedback function. The Comparison of retrieving results by the same queries between Wangda and two other WEB search engines in CERNET(Compass and WebGather) is given out also within the last part of the paper.

【Keywords】Search engine, Information retrieval, RSV, Computer Network